

工作周报总结

2015-05-31

本周工作：

气象项目：

1. 智慧城市、农业数据项目

- 主要实际远程登录到 10.76.6.118 的集群的主节点上，查看了当前的集群的 Spark 的部署情况，简要对已有的一些 python 的脚本进行了运行的测试，对硬件的情况进行了监控处理。
- 目前的硬件的情况，主节点的服务器的内存大约 200G，其余的 10 各左右的节点的内存分布情况不均匀，有的只有不到 1G，大部分是 2G 的内存，spark 的 map reduce 框架限制了每台节点的内存向最低值保持平均，因此后期处理大数据的数据库数据可能还有对内存进行相应的硬件调整。
- Spark 提供的数据集操作类型有很多种，不像 Hadoop 只提供了 Map 和 Reduce 两种操作。比如 map, filter, flatMap, sample, groupByKey, reduceByKey, union, join, cogroup, mapValues, sort, partitionBy 等多种操作类型，他们把这些操作称为 Transformations。同时还提供 Count, collect, reduce, lookup, save 等多种 actions。
- Spark 的中间数据放到内存中，对于迭代运算效率比较高。

2. 其余事情

- 进行了组会论文报告，论文是 VAST 2014 的《Progressive Visual Analytics: User-Driven Visual Exploration of In-Progress Analytics》本文提出了一种渐进式可视分析范式，解决大数据算法的等待时间久的问题，该范式使用户能够在算法的计算过程中对中间结果或部分结果进行可视分析，达到渐进式可视分析的目的。通过论文的报告，我觉得这篇论文是在进行对大数据的可视分析工具开发的一篇非常实际的思想，从系统的可视化 pipeline 到实际细节的处理，非常有针对性。
- 电脑操作系统不小心跪了，用了小两天才恢复~惨~

下周工作：

继续熟悉 Spark 框架的知识，